Ordem de Sensoreamento de Canais em Redes de Rádios Cognitivos Multi-Usuário

André Chaves Mendes[1], **Marcel William Rocha da Silva**[1], **Raphael Melo Guedes**[1], **José Ferreira de Rezende**[1]

[1]Universidade Federal do Rio de Janeiro (UFRJ)
Caixa Postal 68.504 – 21.945-970 – Rio de Janeiro – RJ – Brasil

{andre,marcel,raphael,rezende}@land.ufrj.br

Abstract. *This work investigates the problem of channel sensing order used in a multi-user environment, where each user is able to perform sensing on only one channel at a time. We consider a multichannel cognitive network where the probability of each communication channel being available, and the channel capacity, are not known a priori. Thus, becomes necessary a careful ordering of the sequence of channels that will be sense each time and a tradeoff between maximizing the immediate reward, given by choosing the best sequence, and the refinement of the channel statistics, obtained by exploitation of sub-optimal channels. Therefore, we propose and evaluate an approach using reinforcement learning to search dynamically for the optimal sensing order, comparing its performance with other mechanisms, and the results obtained are superior to the other mechanisms in most of the scenarios.*

Resumo. *Neste trabalho investigamos o problema da escolha da ordem de sensoreamento de canais em um ambiente multi-usuário, onde cada usuário é capaz de realizar o sensoreamento em apenas um canal por vez. Consideramos uma rede de rádios cognitivos multicanal onde a probabilidade de disponibilidade dos canais de comunicação, e sua capacidade, não são conhecidas a priori. Assim, faz-se necessária uma criteriosa ordenação da sequência de canais que será sensoreada a cada vez e um balanceamento entre maximizar a recompensa imediata dada pela escolha da melhor sequência e o refinamento das estatísticas de canal obtidas pela exploração de canais sub-ótimos. Com isso, propomos uma abordagem utilizando aprendizado por reforço para busca dinâmica da ordem de sensoreamento ótima e a avaliamos, comparando o seu desempenho com o de outros mecanismos, obtendo resultados superiores para a maioria dos cenários.*

1. Introdução

Recentes avanços na área da comunicação via rádio e a orientação dada por órgãos regulamentadores, iniciando pelo FCC [FCC 2003], tem reconhecido a ideia do acesso dinâmico ao espectro como uma forma de melhorar o seu próprio uso e, desta forma, fornecer uma solução para a crescente demanda por faixas do espectro de radiofrequências. Esse acesso dinâmico ao espectro apoia-se no advento de um dispositivo de rede reconfigurável capaz de adaptar dinamicamente seus parâmetros e modos de operação às condições do ambiente onde ele se encontra, o chamado *rádio cognitivo* [J. Mitola III and G. Q. Maguire Jr. 1999].

De forma a não causar interferência prejudicial à rede primária, faz-se necessário que os usuários secundários, aqueles dotados do rádio cognitivo, se certifiquem de que não está ocorrendo nenhuma atividade de usuários primários na mesma faixa de frequências antes de qualquer tentativa oportunista de comunicação. Portanto, o sensoreamento do espectro é parte importante no funcionamento desses dispositivos.

Tipicamente, em um cenário de múltiplos canais e usuários secundários dotados de um único transceptor, apenas uma faixa de frequências pode ser sensoreada por vez, devendo essa busca por uma faixa livre ser realizada o mais rapidamente possível. Em sistemas desse tipo, a ordem de sensoreamento, aquela que indica a sequência de canais sensoreados pelos usuários secundários na busca por um canal disponível para uso, é crítica para minimizar esse tempo de busca e acesso a uma faixa livre do espectro e, consequentemente, maximizar a vazão do sistema como um todo.

Quando há conhecimento prévio e preciso das estatísticas dessas faixas de frequências, ou canais, a ordem de sensoreamento *intuitiva*, aquela que segue a ordem decrescente das probabilidades de disponibilidade dos canais é reconhecidamente a ordem ótima [Ho Ting Cheng and Weihua Zhuang 2011]. Entretanto, em muitos cenários práticos, a probabilidade de disponibilidade dos canais não é conhecida previamente, e por isso, a ordem de sensoreamento ótima exige um enorme esforço para ser obtida. Uma primeira abordagem intuitiva para resolver esse problema seria a observação histórica dessa estatística, porém, essa abordagem aumentaria muito o tempo de acesso ao canal devido à variação dessa probabilidade até a estacionariedade e na análise necessária para obter uma estatística precisa.

Basicamente, essa questão recai no problema do compromisso entre maximizar o ganho imediato através da utilização dos canais com melhores disponibilidades conhecidas até o momento ou no refinamento das estatísticas dos canais através da investigação dos canais aparentemente sub-ótimos, que podem ter ganhos maiores. Contudo, a tarefa envolvida não é simples, pois o número possível de sequências ordenadas de canais cresce exponencialmente com o aumento do número de canais, além da escolha de qual canal que será sensoreado a cada vez ser um processo aleatório, mesmo quando seguindo uma determinada ordem de sensoreamento. E nesse caso, tanto a estatística de canal que está sendo investigada quanto o ganho a ser obtido são difíceis de serem quantificadas, e por isso, o processo de aprendizagem envolvido torna-se complexo.

Na solução apresentada em [J. Jia et al. 2008], todos os canais possuem a mesma probabilidade de disponibilidade. Neste caso, o problema de ordem de sensoreamento se reduz ao uso da sequência de canais em ordem decrescente de suas taxas alcançáveis, como em [Ho Ting Cheng and Weihua Zhuang 2011].

O trabalho descrito em [Shu and Krunz 2009] considerou apenas o sensoreamento (e acesso) conforme a ordem crescente dos canais, assumindo limitações de *hardware* e de energia nos rádios cognitivos, e utilizou como métrica a disponibilidade e a qualidade dos canais na sua proposta de uma estratégia de vazão eficiente, embora tenha assumido somente a existência de canais homogêneos.

Em [H. Jiang et al. 2009] é proposto o uso da programação dinâmica para reduzir a complexidade computacional da busca pela ordem de sensoreamento ótima, encontrando uma solução sub-ótima com complexidade de $O(N.2^{N-1})$, para canais com

probabilidade de disponibilidade distintos. Seguindo a mesma linha, os autores em [Han Han et al. 2010] fornecem também uma solução sub-ótima, baseada em árvore de decisão com uma complexidade de $O(N^3)$. Essas duas soluções são comparadas nesse último trabalho, além das sequências aleatória e em ordem decrescente de disponibilidade, denominada como "sequência intuitiva" em [H. Jiang et al. 2009].

Em [Ho Ting Cheng and Weihua Zhuang 2011], os autores propõem o uso da sequência de canais em ordem decrescente de suas taxas alcançáveis, não necessitando assim do conhecimento a priori da atividade dos primários. É demonstrado no trabalho que se a regra de parada utilizada for a do primeiro canal livre, a melhor recompensa dessa sequência é alcançada.

Na sua maioria, esses trabalhos fazem uso da regra de parada tradicional (*traditional stopping rule*) [R. Fan and H. Jiang 2009] para encontrar a melhor sequência de sensoreamento com único usuário secundário e assumem que as estatísticas dos canais são conhecidas. Quando expandida para o caso de múltiplos usuários secundários, a regra de parada tradicional não leva a sequência ótima [R. Fan and H. Jiang 2009]. Além disso, por apresentarem uma alta complexidade computacional com o aumento do número de canais, não é possível embarcá-los nos rádios cognitivos com facilidade.

Com uma outra abordagem, alguns trabalhos se valem da semelhança desse problema com o *problema do bandido de n-braços* [Berry and Fristedt 1985] na busca de uma solução ótima. No trabalho descrito em [L. Lai et al. 2007], os autores aplicaram formulações derivadas das obtidas para o problema do bandido de n-braços em outro trabalho [Auer et al. 2002] para encontrarem soluções para o problema de seleção de canal em redes cognitivas. Os autores em [Liu and Zhao 2010] seguiram a mesma linha, mas com uma abordagem descentralizada, além de apresentarem uma heurística que tende para a solução ótima, quando o tempo de observação dos canais cresce muito. Entretanto, nesses trabalhos assumiu-se que dentro de um tempo fixo de observação, não seria possível sensorear mais de um canal.

No nosso trabalho em [Mendes et al. 2011], investigamos o problema da escolha da ordem de sensoreamento dos canais para apenas um usuário em um ambiente multicanal e propomos uma abordagem utilizando um método de baixa complexidade com um algoritmo baseado em uma máquina de aprendizagem por reforço (*reinforcement learning*) para busca dinâmica da ordem de sensoreamento ótima e realizamos sua avaliação e comparação com outros mecanismos, inclusive com a sequência ótima obtida por força-bruta, obtendo-se resultados muito próximos do ótimo.

Como continuidade, neste trabalho investigamos o problema da escolha da ordem de sensoreamento dos canais em ambiente multi-usuário, e propomos um mecanismo que realiza uma criteriosa ordenação da sequência de canais que será sensoreada a cada vez, fazendo uso de uma máquina de aprendizagem por reforço, seguindo o modelo de recompensas que se utiliza da teoria da parada ótima [Chow et al. 1971].

Nesta solução, como na que propusemos em [Mendes et al. 2011], não é necessário o conhecimento prévio dos momentos das variáveis de disponibilidade e capacidade dos canais, possibilitando uma adaptação dinâmica às variações desses momentos. Além disso, essa solução também possui complexidade computacional baixa, o que a torna atrativa para ser embarcada no rádio cognitivo, tendo sido implementada num simu-

lador próprio e avaliada em comparação a outros mecanismos mais simples de ordenação.

No restante deste artigo, a seção seguinte descreve o modelo do sistema utilizado. A Seção 3 apresenta a nossa proposta para o aprendizado dinâmico da ordem de sensoreamento ótima. A Seção 4 descreve o ambiente de simulação e mostra os resultados obtidos. Finalmente, a Seção 5 conclui o artigo e enumera trabalhos futuros.

2. Modelo do Sistema

Nesta seção, descreveremos o modelo do sistema adotado para o desenvolvimento e implementação da nossa proposta. Esta modelagem permite determinar a sequência de sensoreamento ótima através da aplicação da teoria da parada ótima (*optimal stopping*) [Chow et al. 1971]. Desta forma, o objetivo é fornecer uma decisão sobre o momento para finalizar o sensoreamento de novos canais de tal forma que a recompensa obtida na escolha de um canal seja maximizada. Essa teoria permite então definir a regra de parada que maximiza a recompensa.

Considere uma rede com múltiplos usuários secundários e um número finito de canais, N. Cada usuário é equipado com um transceptor para a troca de dados que é sintonizado em um dos N canais da rede. Além disso, cada usuário possui um modelo de funcionamento cujo tempo de acesso, que chamaremos *slot*, destinado à observação do canal e a transmissão de dados é constante com duração T. Em cada *slot*, cada canal i possui a probabilidade p_i de estar livre da atividade de usuários primários, i. e., de apresentar-se livre ou ocupado. Assume-se que esta probabilidade é independente do seu estado prévio e do estado de outros canais, dentro de cada *slot*, e i.i.d. entre *slots* [L. Lai et al. 2007, H. Jiang et al. 2009].

Consideramos também que, devido ao efeitos de desvanecimento nos canais, a relação sinal-ruído (SNR) obtida em um canal varia aleatoriamente entre *slots*. Assumimos que essa SNR aleatória é i.i.d entre os *slots* e os diferentes canais, e que é regida por uma distribuição arbitrária. Se o usuário secundário decidir transmitir em um canal considerado livre, c_i, a taxa de transmissão obtida será função da SNR momentânea desse usuário nesse canal. Esta função, $F(SNR_i)$, mapeia de forma monotônica e crescente a SNR do canal c_i na taxa de transmissão que será obtida neste canal.

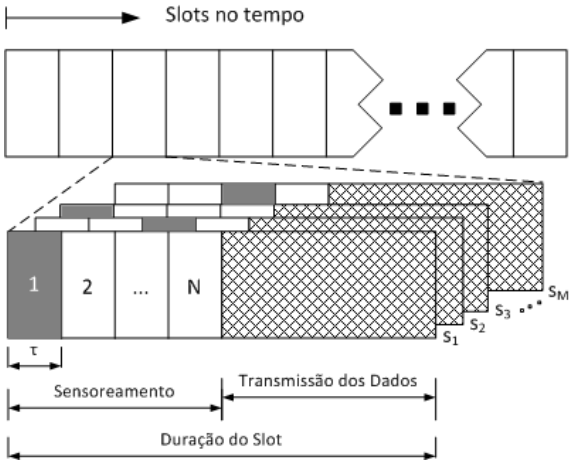


Figura 1. Processo de sensoreamento dos canais em um *slot*.

A Figura 1 exemplifica a atividade de um usuário secundário em um *slot*, o qual possui duas fases: uma fase de sensoreamento e uma fase de transmissão de dados. Antes de decidir utilizar um canal em um determinado *slot*, o usuário secundário deve realizar o sensoreamento do canal com a finalidade de determinar se nele existem usuários primários em atividade. O valor fixo τ corresponde ao tempo necessário para o sensoreamento de cada canal de tal forma a minimizar as probabilidades de não detecção do usuário primário e de alarmes falsos. No modelo, assume-se que o processo de sensoreamento é uma tarefa precisa e isenta de erros.

Como não existe conhecimento prévio a respeito dos estados dos canais, durante a fase de sensoreamento, cada usuário secundário realiza o sensoreamento sequencial dos N canais. Portanto, poderão ser realizados no máximo N sensoreamentos de canal por *slot*. Chamaremos de k o número máximo de sensoreamentos por *slot*, sendo esse valor igual ao $\min(N, \lfloor(\frac{T}{\tau})\rfloor)$. Com isso, a quantidade de sequências possíveis é dada por $M = \begin{pmatrix} N \\ k \end{pmatrix} k$!. Assim, cada usuário secundário segue uma ordem pré-estabelecida dada pela sequência $s_n = \{o_1, o_2, ..., o_N\}, n < M$, correspondente a uma permutação dos N canais disponíveis, até decidir em qual canal parar.

Ao estabelecerem o canal que irão utilizar, os usuários secundários realizam o controle de acesso ao meio da seguinte forma: inicialmente, um usuário secundário que deseja acessar o meio efetua o sensoreamento do canal para verificar se este está sendo utilizado por um usuário primário e, caso não esteja, pode haver disputa pelo canal com outros usuários secundários ou colisão, conforme o valor instantâneo de uma variável aleatória em comparação com um limiar dado pela probabilidade de colisão, segundo a fórmula $\mathrm{Pr}_{\text{colisão}} = 1 - (1 - \frac{1}{W_m})^{(U-1)}$ [Vu and Sakurai 2006], onde o número de usuários secundários é dado por U e W_m é o tamanho médio da janela de contenção no padrão IEEE 802.11 [802.11b 1999].

Se houver disputa, esses usuários secundários aguardam durante um período de tempo, cuja duração é sorteada dentro de um intervalo determinado, e ao final desse período retornam ao sensoreamento do canal, para garantir que esteja livre, e depois o utilizam efetivamente ou perdem o *slot*. Nesse caso, a recompensa é coletada apenas pelo vencedor da disputa, enquanto que se há colisão, nenhuma recompensa é coletada.

A eficiência no uso de uma sequência de sensoreamento está relacionada com o tempo gasto no sensoreamento dos canais até que seja encontrado um canal adequado para ser utilizado e com a taxa de transmissão momentânea que pode ser obtida nesse canal. Desta forma, caso o i-ésimo canal a ser sensoreado, c_i, seja considerado como livre, a recompensa coletada é a taxa de transmissão efetiva obtida pelo usuário secundário no uso deste canal durante o tempo remanescente do *slot* sendo dada por $r_i = e_i \times F(SNR_i)$, onde e_i é a efetividade da transmissão, calculada pela fórmula $e_i = 1 - \frac{i\tau}{T}$.

3. Ordem de Sensoreamento Dinâmica Utilizando Aprendizado por Reforço

O mecanismo proposto neste trabalho utiliza aprendizado por reforço para determinar de maneira dinâmica uma ordem de sensoreamento a ser utilizada em cada *slot*. Uma das vantagens desse mecanismo é que não é necessário nenhum conhecimento prévio a respeito da probabilidade de cada canal estar disponível, nem da qualidade estimada de cada canal por meio de suas SNRs médias. Outra vantagem importante desta proposta

é quanto a sua adaptabilidade às mudanças de características dos canais garantida pelo aprendizado com as tomadas de ação. Logo, o mecanismo torna-se imune às possíveis mudanças nas probabilidades de disponibilidade dos canais, que podem ocorrer devido a mudanças nos padrões de atividade dos usuários primários, e às possíveis mudanças na qualidade dos canais (SNRs médias), que podem ocorrer devido à mobilidade e aos efeitos de desvanecimento de larga escala.

Nas subseções seguintes, apresentaremos os conceitos básicos sobre a técnica de aprendizado por reforço e a nossa proposta, baseada nessa técnica, para a busca da ordem de sensoreamento ótima.

3.1. Aprendizado por Reforço

Aprendizagem por reforço (*reinforcement learning*) [Sutton and Barto 1998] é um tipo de máquina de aprendizagem que preocupa-se com a maneira com que um agente escolhe ações a serem realizadas, de acordo com informações de causa e efeito obtidas do ambiente. Resumidamente, nesse método temos agentes que ao inspecionarem um estado, em um espaço de estados, realizam alguma ação que se reflete em uma recompensa. A partir do valor da recompensa coletada, o agente aprende a qualidade da ação escolhida. O problema consiste em escolher ações que maximizem o total de recompensas recebidas pelo agente.

Entre as diversas técnicas de aprendizagem existentes [Sutton and Barto 1998], baseamos este trabalho na técnica *Q-learning* por sua simplicidade, e por este motivo a apresentaremos em mais detalhes. Nesta técnica cada agente mantém uma *Q-table*, que é uma matriz com $|S| \times |A|$ entradas, onde as linhas representam os estados e as colunas indicam as ações. Os elementos dessa matriz são chamados *Q-value's*, $Q_t(s, a)$, que são valores atualizados através da coleta de recompensas, $r_t(s, a)$, sempre que um agente realiza uma ação a em um estado s, em um dado instante t.

O *Q-value* estima o nível de recompensa para o par estado-ação. Assim mudanças nos *Q-value's* levam a mudanças nas decisões de que ações devem ser tomadas pelos agentes. A cada instante de decisão t, o agente observa seu estado atual (linha) e escolhe uma ação (coluna) em sua *Q-table*. Posteriormente à execução de uma ação, o agente recebe uma recompensa r_t relativa a esta ação realizada. A partir da recompensa obtida, o agente atualiza a respectiva entrada na *Q-table* conforme a Equação 1, onde o valor de cada variável corresponde ao instante t, exceto quando explicitado em contrário:

$$\begin{aligned} &Q_{t+1}(s,a) = \\ &Q(s,a) + \alpha[r(s,a) + \gamma max_a Q(s_{t+1},a) - Q(s,a)] \end{aligned} \tag{1}$$

Na Equação 1, α é chamado de parâmetro de aprendizagem, e γ é chamado de fator de desconto. Maiores valores de α indicam maior importância para a experiência recente em relação ao histórico. Valores maiores de γ indicam que o agente baseia-se mais na recompensa futura que na recompensa imediata [K. A. Yau et al. 2010].

O *Q-learning* inicia com a *Q-table* zerada e a cada instante o agente seleciona uma ação baseada em uma estratégia de exploração. Uma estratégia comumente usada é a ε*-greedy* [Sutton and Barto 1998], onde o agente utiliza a probabilidade dada por ε para

decidir entre a exploração (*exploitation*) da *Q-table* ou a investigação (*exploration*) de estados aleatoriamente. Dado que ε é normalmente pequeno, na maioria dos casos, o agente seleciona de forma gananciosa a ação que satisfaça $max_a Q(s, a)$, ou seja, aquela que historicamente oferece a maior recompensa. No entanto, ocasionalmente, o agente seleciona uma ação aleatória com uma probabilidade ε. Esta estratégia tenta fazer com que todas as ações, e seus efeitos, sejam experimentados [Jelle R. Kok and Nikos Vlassis 2006].

3.2. Proposta

Neste trabalho cada usuário na rede secundária é modelado como um "agente" de aprendizagem. Em um determinado instante, esse "agente" observa somente o seu próprio "ambiente" de aprendizagem devido à sua capacidade de sensoreamento limitada e no próximo instante, esse "agente" coleta sua recompensa local seguindo sua decisão de melhor ação. Com isso, o mecanismo de aprendizagem obtém conhecimento do "ambiente" de aprendizagem observando as consequências das ações tomadas anteriormente [Sutton and Barto 1998]. E com o tempo, esse "agente" aprende a melhor ação que leva à maximização da recompensa.

Um dos maiores desafios encontrados no emprego da ferramenta de aprendizado por reforço no problema da escolha da ordem de sensoreamento foi a modelagem dos estados e ações. Uma modelagem descuidada pode gerar um modelo com muitos estados e/ou muitas ações, o que tornaria lenta a convergência do processo de exploração. No nosso modelo, definimos o estado como o par ordenado formado pela posição na ordem de sensoreamento, o_k, e o canal que é sensoreado naquela posição, c_i. As ações possíveis de serem tomadas por um usuário secundário a partir de um estado (o_k, c_i) correspondem a escolher o canal que será sensoreado na próxima posição da ordem de sensoreamento, o_{k+1}.

Com isso, a *Q-table* será uma matriz de dimensões $N^2 \times N$ ($estados \times$ ações). Repare que essa modelagem faz com que não exista um estado ótimo a ser alcançado, mas sim uma sequência de ações que maximizam a recompensa imediata a cada estado, criando uma ordem de sensoreamento dinâmica.

Algumas restrições devem ser levadas em consideração no momento das tomadas de ação e atualização da *Q-table*. Uma delas é que uma ação tomada no estado $(o_k, *)$, com $1 \leq k \leq (N-1)$, sempre leva a um estado onde a posição na ordem de sensoreamento é o_{k+1}. No estado onde a posição é $(o_N, *)$, que representa o último canal da ordem de sensoreamento, as ações indicam o primeiro canal a ser sensoreado no próximo *slot*, ou seja, leva a um estado onde a posição na ordem de sensoreamento é a $(o_1, *)$. Quando o usuário secundário decide usar um canal c_i na posição o_k, o sensoreamento nesse *slot* é finalizado. Nesse caso, o primeiro canal a ser sensoreado no próximo *slot* será determinado pela melhor ação no estado (o_N, c_i).

Outra restrição importante é com relação a impedir o retorno a um canal sensoreado previamente (*recall*). Para isso, é necessário armazenar os canais já sensoreados no *slot* corrente. Desta forma, antes de tomar uma ação, o usuário secundário deve eliminar, das ações possíveis, os canais já sensoreados.

Uma parte importante do modelo diz respeito à recompensa obtida em cada escolha de canal, o qual será utilizado para atualizar a *Q-table*, de acordo com o modelo de recompensa apresentado na Seção 2. *Individualmente*, quando o canal c_i é sensoreado

como *livre* na posição o_k da ordem de sensoreamento, a recompensa obtida r_t será dada pela taxa de transmissão efetiva obtida pelo uso daquele canal durante o tempo remanescente do *slot*, $e_k \times F(SNR_i)$. No caso em que o canal é detectado como *ocupado*, é necessário que exista alguma penalização para reduzir o *Q-value* referente àquela ação. Desta forma, introduzimos o parâmetro δ, que assume valores no intervalo $[0, 1]$ e multiplica o *Q-value* atual referente àquela ação. Assim, garante-se que quando uma ação leva a um canal ocupado, o *Q-value* referente àquela ação será reduzido. Essa estratégia faz com que o *Q-value* represente não apenas a taxa de transmissão efetiva, mas também a disponibilidade dos canais.

Assim, os valores da *Q-table* são atualizados conforme a Equação 2, onde o valor de cada variável corresponde ao instante t, exceto quando explicitado em contrário:

$$Q_{t+1}(s,a) = \begin{cases} (1-\alpha) \times Q(s,a) + \alpha \times [r(s,a) + \gamma max_a Q(s_{t+1}, a)] & se \text{ canal livre} \\ \delta \times Q(s,a) & se \text{ canal ocupado} \end{cases} \quad (2)$$

O funcionamento do mecanismo é descrito pelo Algoritmo 1. No início, todos os pares estado-ação da *Q-table* são completados com zeros. Realizada esta fase de inicialização, começa a fase de aprendizado, que é repetida durante todo o período de funcionamento do mecanismo. Nesta fase, toma-se a decisão entre *investigação*, onde uma ação é escolhida aleatoriamente, e *exploração*, onde a melhor ação é escolhida, baseando-se na *Q-table*. Após a execução da ação, o mecanismo torna-se capaz de calcular a recompensa obtida e atualizar o correspondente *Q-value*.

Uma característica importante da nossa proposta diz respeito ao *uso dos canais sensoreados como livres*. De acordo com o modelo apresentado na Seção 2, o objetivo é fornecer uma decisão sobre o momento para finalizar o sensoreamento de novos canais de tal forma que a recompensa obtida na escolha de um canal seja maximizada. Isso indica que nem sempre será vantajoso utilizar o primeiro canal livre encontrado.

De forma similar, a nossa proposta também utiliza um critério de parada que consiste em comparar a recompensa atual, r_t, com o melhor *Q-value* das ações possíveis a partir daquele estado. Assim, é possível estimar se a recompensa do canal livre atual é superior à recompensa esperada da melhor ação existente. Repare que mesmo no caso onde o canal livre não é utilizado, o *Q-value* referente àquela ação também é atualizado.

4. Resultados Numéricos

Para avaliar o comportamento do mecanismo de aprendizado por reforço na solução do problema da ordem de sensoreamento, desenvolvemos um simulador utilizando a linguagem Tcl [John Ousterhout 1988] que emula o funcionamento de uma rede secundária, onde cada um de seus usuários utiliza sequências de sensoreamento individuais.

Nesse simulador, as seguintes ordens de sensoreamento foram avaliadas: a sequência dinâmica dos canais fornecida pela nossa proposta (`RL`), a sequência de canais na ordem decrescente de suas probabilidades de disponibilidade (`Prob`), a

while *! fim_do_sensoreamento* **do**
  $x = Uniforme(0, 1)$;
  if *(*$x < \varepsilon$*)* **then**
    $a = Uniforme(c_1, c_N)$;
  else
    /* $s = s_t$ é o estado atual */
    $a = argmax_s(Q(s, *))$;
  end
  if *(canal livre)* **then**
    /* canal c_a correspondente à ação a */
    calcula recompensa $r_t(s, a)$;
    $Q_{t+1}(s, a) \leftarrow (1 - \alpha) \times Q_t(s, a) + \alpha \times [r_t(s, a) + \gamma max_a Q_t(s_{t+1}, a)]$;
    if $r_t(s, a) > max_a Q(s, a)$ **then**
      /* usa canal c_a */
      fim_do_sensoreamento = 1;
    else
      /* não usa canal c_a */
      continua sensoreamento;
    end
  else
    /* canal c_a ocupado */
    $Q_{t+1}(s, a) \leftarrow [\delta \times Q_t(s, a)]$;
    continua sensoreamento;
  end
  $s_t = s_{t+1}$;
end

Algoritmo 1: Mecanismo proposto baseado em aprendizado por reforço.

sequência dada pela ordem decrescente das capacidades médias de cada canal (`Cap`) [Ho Ting Cheng and Weihua Zhuang 2011] e a sequência aleatória dos canais (`RND`).

Vale ressaltar que todas as sequências avaliadas, com exceção da sequência `RL`, são estáticas, ou seja, não mudam durante toda a simulação. No caso da sequência `RL`, devido ao próprio aprendizado por reforço, a sequência pode variar durante a simulação. Além disso, todas as sequências, exceto na `RL` e `RND`, assumem o conhecimento a priori das capacidades médias de cada canal e/ou de suas probabilidades de disponibilidade.

4.1. Modelo de Simulação

No início de cada rodada de simulação, sorteamos o valor da capacidade média e as probabilidades de disponibilidade de cada canal i, $i \in \{1, ..., N\}$. A capacidade média é sorteada seguindo uma distribuição uniforme dentro do intervalo $[0.1 \times CAPMAX, CAPMAX]$, onde $CAPMAX$ é a *capacidade média máxima* dos canais. A probabilidade de disponibilidade de cada canal é sorteada uniformemente dentro do intervalo [0,1].

Numa rodada de simulação, a cada *slot* T, o estado do canal (livre ou ocupado) é sorteado de acordo com a sua probabilidade de disponibilidade utilizando-se uma distribuição uniforme. Além disso, a capacidade instantânea de cada canal é sorteada

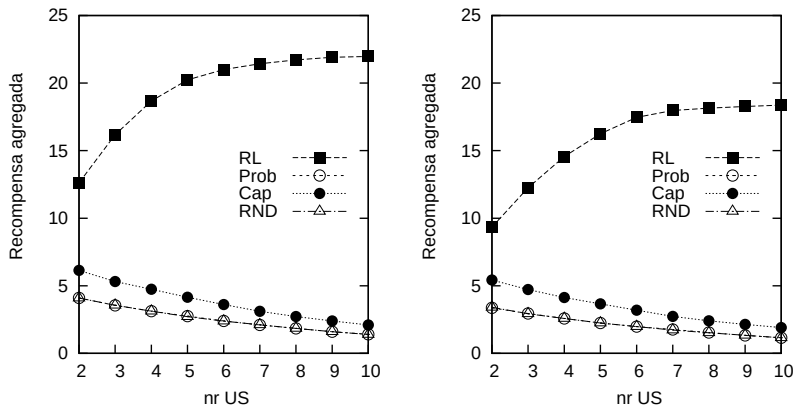
utilizando-se uma distribuição uniforme dentro do intervalo $[0, 2 \times CAPMEDIA]$, onde $CAPMEDIA$ é a *capacidade média* de cada canal.

A cada *slot* T, o simulador calcula a recompensa obtida por cada uma das sequências implementadas usando-se os mesmos estados e as mesmas capacidades instantâneas dos canais (critério de justiça). A recompensa em cada *slot* corresponde à taxa de transmissão efetiva (Seção 2) e a recompensa global é dada pela soma das recompensas individuais. Um rodada de simulação consiste na execução de X *slots*. Ao final de cada rodada, o simulador fornece a recompensa média obtida por cada uma das sequências em todos os X *slots* para cada usuário secundário.

4.2. Resultados

Foram realizadas simulações com 1.000 *slots* cada, utilizando-se um número de canais variando de 3 a 10. O valor de W_m, referente à probabilidade de colisão, foi configurado em 8 [Vu and Sakurai 2006]. Adotamos o valor inicial de 0.7 para ε baseado no compromisso entre número de estados e os resultados obtidos, favorecendo a exploração de um número maior de estados dentro do período de aprendizado, correspondente a 20% do número total de *slots*, passando para 0.1 ao final desse período. O parâmetro α do `RL` foi configurado em 0.9 e o parâmetro γ em 0.1, beneficiando a experiência recente face ao histórico. O tamanho do *slot* T é um múltiplo inteiro do tempo necessário para sensorear um canal (τ). Foram feitas 30 rodadas de simulação para cada conjunto de parâmetros. Em todas as simulações, $CAPMAX$ foi configurada com valor fixo de 10 e o tamanho do *slot* foi configurado como variável, com valor igual ao dobro do número de canais utilizados na simulação multiplicado por τ. Em todos os resultados, apresentamos a soma das médias das recompensas coletadas a cada rodada, por cada usuário secundário.

O estado do canal (livre ou ocupado) foi estabelecido seguindo um modelo *on-off* exponencial e, durante as simulações, o canal permaneceu no estado ocupado (*OFF*) por um tempo seguindo uma distribuição exponencial de média igual a t_{OFF}. Desta forma, o t_{ON} pode ser obtido por $t_{ON} = \frac{(1-u) \times t_{OFF}}{u}$, onde u é a *utilização do canal* pelos usuários primários.



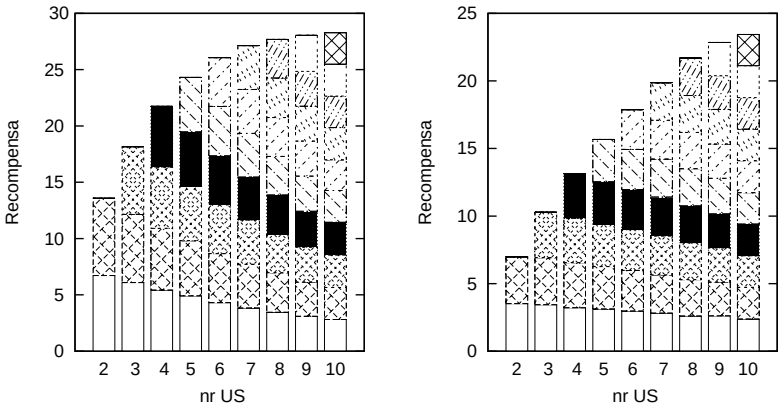
(a) Utilização = 10% (b) Utilização = 90%

Figura 2. Resultados para 7 canais da nossa proposta (`RL`), da ordem decrescente de suas probabilidades de disponibilidade (`Prob`), da ordem decrescente de capacidades médias (`Cap`) e da ordem aleatória (`RND`).

Nesta primeira parte, apresentamos os resultados para 7 canais para utilização dos canais pelos usuários primários igual a 10% (Figura 2(a)) e para 90% (Figura 2(b)). A comparação do desempenho das sequências neste gráfico mostra que a nossa proposta, `RL`, apresenta resultados melhores. O desempenho das outras sequências é desfavorável, pois nenhuma delas utiliza regras de parada baseadas na predição do desempenho estimado de se continuar sensoreando os próximos canais da sequência, ou seja, nestas outras soluções o primeiro canal sensoreado como livre sempre é utilizado. Desta forma, o `RL`, que utiliza as experiências passadas armazenadas na *Q-table*, consegue determinar de maneira eficiente se é vantajoso utilizar um determinado canal sensoreado como livre.

Uma observação interessante a respeito das curvas das Figuras 2(a) e 2(b) é que o desempenho das sequências `Prob`, muito próximo ao das sequências aleatórias `RND`, é inferior ao desempenho das sequências `Cap`. Isso indica que neste cenário a diferenciação entre as capacidades médias dos canais ($CAPMEDIA$) é mais importante do que a diferenciação entre suas probabilidades de disponibilidade. Com isso, é melhor ordenar os canais pela ordem decrescente de suas capacidades médias, pois aumenta-se a probabilidade de o primeiro canal sensoreado como livre ser um canal de maior capacidade.

Outro detalhe é que a curva referente à nossa proposta, `RL`, apresenta um crescimento menor conforme aumenta o número de usuários secundários, indicando que há um limiar de saturação da capacidade disponível da rede secundária conforme o número de canais disponíveis.

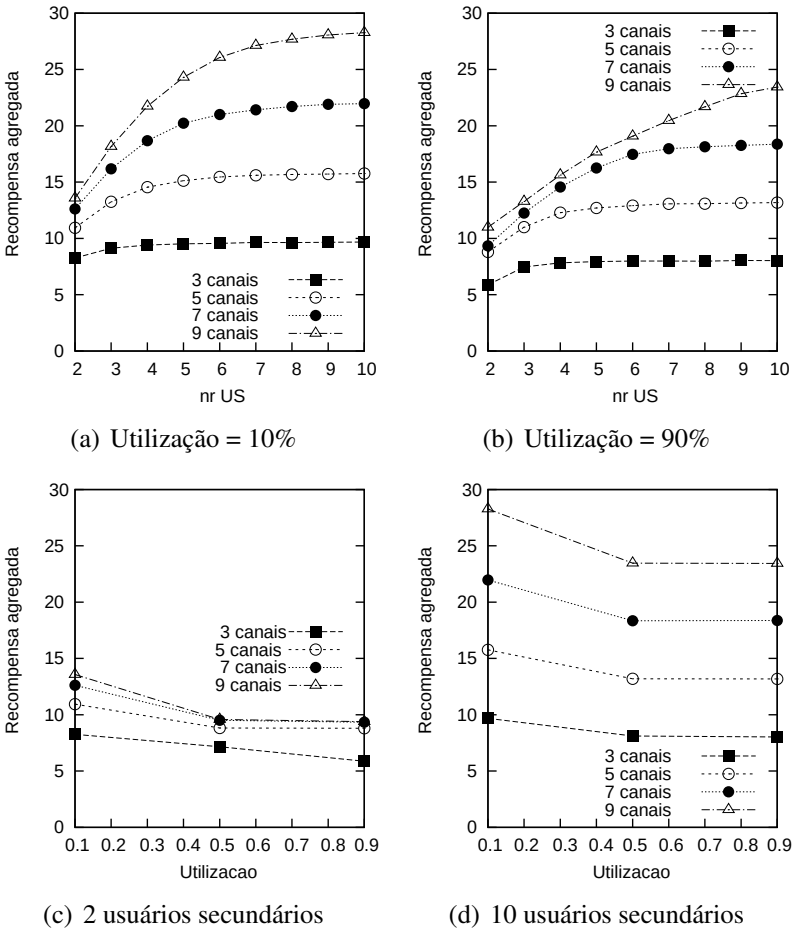


(a) Utilização = 10% (b) Utilização = 90%

Figura 3. Medida de justiça entre os usuários secundários, para utilização de 10% e 90%.

Em seguida, verificamos a justiça entre os usuários secundários que seguiram a sequência dinâmica individual fornecida pela nossa proposta (`RL`) (Figura 3). Embora a existência de múltiplos usuários pudesse colaborar para um desequilíbrio neste critério, uma vez que um dos usuários secundários poderia tornar-se ganancioso, isso não foi constatado nem para utilização dos canais pelos usuários primários igual a 10% (Figura 3(a)) nem para 90% (Figura 3(b)). Foram feitas verificações para o número de canais variando entre 3 e 10 e observado que este comportamento ganancioso não se estabeleceu, pois a oferta das oportunidades de canais é igual para todos os usuários secundários, que possuem também a mesma probabilidade de acesso ao meio.

Nesta segunda parte, tratamos dos possíveis impactos que a nossa proposta, RL, poderia sofrer com a variação do número de canais disponíveis, com a variação da utilização dos canais pelos usuários primários e com a competição entre os usuários secundários, dado pela medida das colisões na rede secundária.



(a) Utilização = 10%

(b) Utilização = 90%

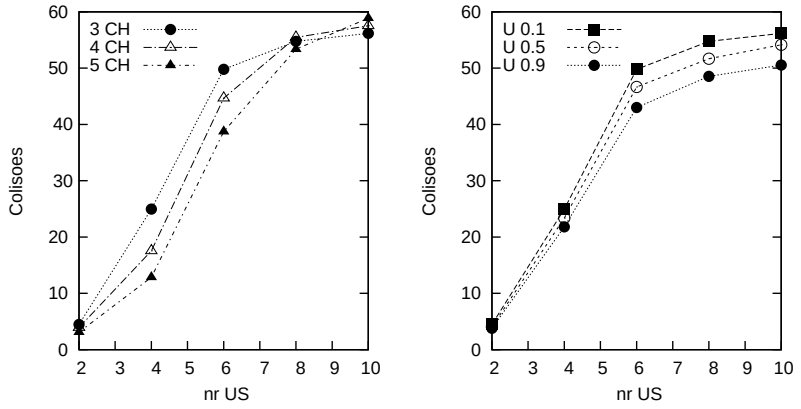
(c) 2 usuários secundários

(d) 10 usuários secundários

Figura 4. Impacto da variação da quantidade de canais e do parâmetro utilização no método baseado no RL, para utilização de 10% e 90% e para 2 e 10 usuários secundários.

Iniciamos variando a quantidade de canais e observamos que com a probabilidade do canal estar indisponível pelo acesso de usuários primários pequena (Figura 4(a)) o aumento do número de canais proporcionalmente ao número de usuários traz aumento de recompensa. Por outro lado, aumentando a probabilidade de indisponibilidade do canal (Figura 4(b)), observamos que há um limiar máximo de recompensa que pode ser atingida, onde mesmo que se aumente o número de canais, a recompensa não acompanha na mesma proporção.

Variando a utilização dos canais pelos usuários primários, observamos que com poucos usuários secundários (Figura 4(c)), a recompensa obtida varia pouco com a variação da utilização e dos canais. Quando aumentamos o número de usuários secundários (Figura 4(d)), observamos um aumento da recompensa, embora possamos notar que a variação no parâmetro utilização não impacta muito, mantendo um crescimento percentual da recompensa semelhante com o aumento do número de canais disponíveis.



(a) Utilização = 10% (b) 3 canais

Figura 5. Curvas do número de colisões, para utilização de 10%, conforme os canais disponíveis, e para 3 canais disponíveis, conforme a utilização.

Na Figura 5, analisamos as colisões na rede secundária e observamos que para valor de utilização igual a 10% (Figura 5(a)), o número de colisões aumenta com o aumento do número de usuários na rede secundária, evidenciando um aumento na competição pelo uso do mesmo canal simultaneamente. Porém, também se observa que o número de colisões é menor para um número maior de canais disponíveis, o que era esperado, pois houve aumento das oportunidades com o aumento do número de canais. Outro indicador da disputa na rede secundária pode ser observado na Figura 5(b), onde a curva superior é a de menor valor do parâmetro utilização, indicando maior disponibilidade dos canais.

5. Conclusões

Neste trabalho investigamos o problema da escolha da ordem de sensoreamento de canais em um ambiente multi-usuário, onde cada usuário é capaz de realizar o sensoreamento em apenas um canal por vez a fim de detectar oportunidades de uso. A ordem de sensoreamento indica a sequência de canais sensoreados pelos usuários secundários na busca por um canal disponível para uso. Nesses casos, a ordem utilizada para sensorear os canais pode ter grande impacto no desempenho.

Na nossa abordagem, consideramos uma rede de rádios cognitivos multicanal onde os canais não estão sempre disponíveis, devido as atividades dos primários e propomos uma solução de baixa complexidade que utiliza uma máquina de aprendizado por reforço baseada na técnica *Q-learning*. Essa solução não requer o conhecimento prévio das probabilidades de disponibilidade e da capacidade média esperada em cada canal, podendo se adaptar dinamicamente às variações dessas características. Além disso, ao possuir complexidade computacional baixa, essa solução torna-se também atrativa para ser embarcada em rádios cognitivos.

Os resultados das simulações mostram que o mecanismo proposto obtém desempenho superior aos outros tipos de ordenamento que foram avaliados quando variamos o número de usuários secundários, para diferentes valores de utilização dos canais pelos usuários primários. Outra observação importante mostra a inexistência de usuários secundários gananciosos e que, ao variarmos a quantidade de canais com alta probabilidade de indisponibilidade, há um limiar máximo de recompensa que pode ser atingido, onde

mesmo que se aumente o número de canais, a recompensa não acompanha na mesma proporção.

Como trabalhos futuros, pretendemos estender a avaliação para cenários onde as probabilidades de disponibilidade e as capacidades médias dos canais variam dentro do tempo de *slot* e considerarmos na avaliação a possibilidade de erros no sensoreamento, verificando como o mecanismo proposto reage.

Referências

802.11b (1999). Wireless LAN MAC and PHY Specifications: Higher-Speed Physical Layer Extension in the 2.4GHz Band. IEEE Standard.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite time analysis of the multi-armed bandit problem. *Machine Learning*, 47.

Berry, D. and Fristedt, B. (1985). *Bandit problems: Sequential Allocation of Experiments*. Chapman e Hall.

Chow, Y. S., Robbins, H., and Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin Company.

FCC (2003). FCC-03-322 - NOTICE OF PROPOSED RULE MAKING AND ORDER. Technical report, Federal Communications Commission.

H. Jiang, L. Lai, R. Fan, and H. V. Poor (2009). Optimal Selection of Channel Sensing Order in Cognitive Radio. *IEEE Transactions in Wireless Communications*.

Han Han, Jin-long Wang, Qi-hui Wu, and Yu-zhen Huang (2010). Optimal Wideband Spectrum Sensing Order Based on Decision-making Tree in Cognitive Radio. *International Conference on Wireless Communications and Signal Processing (WCSP)*.

Ho Ting Cheng and Weihua Zhuang (2011). Simple Channel Sensing Order in Cognitive Radio Networks. *IEEE Journal on Selected Areas in Communications*.

J. Jia, Q.Z., and X. Shen (2008). HC-MAC: a Hardware Constrained Cognitive MAC for Efficient Spectrum Management. *IEEE Journal on Selected Areas in Communications*.

J. Mitola III and G. Q. Maguire Jr. (1999). Cognitive Radio: Making Software Radio more Personal. *IEEE Personal Communications*, 6(4):13–18.

Jelle R. Kok and Nikos Vlassis (2006). Collaborative Multiagent Reinforcement Learning by Payoff Propagation. *J. Mach. Learn. Res.*, 7:1789–1828.

John Ousterhout (1988). Tcl - Tool Command Language. http://www.stanford.edu/ ouster/cgi-bin/tclHistory.php.

K. A. Yau, P. Komisarczuk, and P. D. Teal (2010). Enhancing Network Performance in Distributed Cognitive Radio Networks using Single-agent and Multi-agent Reinforcement Learning. In *IEEE Conference on Local Computer Networks (LCN)*.

L. Lai, H. El Gamal, H. Jiang, and H. V. Poor (2007). Cognitive Medium Access: Exploration, Exploitation and Competition. *IEEE Transactions on Networking*.

Liu, K. and Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*.

Mendes, A. C., Augusto, C. H. P., Silva, M. W. R. d., Guedes, R. M., and Rezende, J. F. d. (2011). Seleção da Ordem de Sensoreamento de Canais em uma Rede Cognitiva Oportunista. In *I Workshop de Redes de Acesso de Banda Larga - WRA'11*.

R. Fan and H. Jiang (2009). Channel sensing order setting in cognitive radio networks: a two user case. *IEEE Transactions on Vehicular Technology*.

Shu, T. and Krunz, M. (2009). Throughput efficient sequential channel sensing and probing in cognitive radio networks under sensing errors. In *MobiCom*.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: an Introduction*. MP.

Vu, H. L. and Sakurai, T. (2006). Collision probability in saturated ieee 802.11 networks. In *Australian Telecommunication Networks and Applications Conference*.